Corpus Construction for the BioCreative IV GO Task

Kimberly Van Auken[1], Mary L. Schaeffer[2], Peter McQuilton[3], Stanley J. F. Laulederkind[4], Donghui Li[5], Shur-Jen Wang[4], G. Thomas Hayman[4], Susan Tweedie[3], Cecilia N. Arighi[6], James Done[1], Hans-Michael Müller[1], Paul W. Sternberg[1,7], Yuqing Mao[8], Chih-Hsuan Wei[8] and Zhiyong Lu[8,*]

[1]Division of Biology, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA
[2]USDA-ARS Plant Genetics Research Unit and Division of Plant Sciences, Department of Agronomy, University of Missouri, Columbia, MO 65211, USA
[3]FlyBase, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK
[4]Rat Genome Database, Human and Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA
[5]Department of Plant Biology, Carnegie Institution for Science, 260 Panama Street, Stanford, CA 94305, USA
[6]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711, USA
[7]Howard Hughes Medical Institute, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA
[8]National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, MD 20894, USA

*Corresponding author. Tel: 301-594-7089 Email: zhiyong.lu@nih.gov

Abstract

Gene function curation via Gene Ontology (GO) annotation is a common task among Model Organism Database (MOD) groups. Due to its manual nature, this task is time-consuming and labor-intensive, and thus considered one of the bottlenecks in literature curation. There have been many previous attempts of automatic identification of GO terms and associated information from full text. However, few systems have delivered an accuracy that is comparable to human annotators. One recognized challenge in developing such systems is the lack of marked passage-level evidence text that provides the basis for making GO annotations. To this end, we aim to create a corpus that includes the GO evidence text along with the three essential elements of GO annotations: 1) a gene or gene product, 2) a GO term and 3) a GO evidence code. To ensure our results are consistent with real-life GO annotation data, we recruited a team of eight professional GO curators from the biocuration community, and asked them to follow their routine GO annotation protocols. With the aid of a web-based annotation tool, our annotators marked up nearly 4,000 unique text passages in 200 full-text articles where on average each unique GO term is annotated with four different evidence text passages. Further, our corpus analysis shows that most of the evidence text occurs in the body of the article while only as little as 12% appears in the abstracts. This result demonstrates the necessity of text mining of full text for finding GO terms. Through its use as the official data set for the BioCreative IV GO (BC4GO) task, we expect our unique BC4GO corpus to become a valuable resource for the BioNLP research community.

Introduction

The Gene Ontology (GO) (http://www.geneontology.org) is a controlled vocabulary for standardizing the description of gene and gene product attributes across species and databases (1). Currently, there are about 40,000 GO terms that are organized in a hierarchical manner under three GO sub-categories: molecular function, biological process and cellular component. Since its inception, GO terms have been used in over 126 million annotations to over 9 million gene products (2). The accumulated GO annotations have been shown to be increasingly important in an array of different areas of biological research such as high-throughput omics data analysis and the study of developmental biology (3-5).

Among the 126 million GO annotations, most are derived from automated techniques such as mapping of GO terms to protein domains, motifs (InterPro2GO) (6) or corresponding concepts in one of the controlled vocabularies by UniProt (7); only a very small portion (1.1 million) are derived from manual curation of published experimental results in the biomedical literature (8). While the former approach is efficient in assigning higher-level GO terms, the latter provides more reliable and detailed GO annotations that are critical for the kinds of analyses mentioned above. Generally speaking, the manual GO annotation process first involves the retrieval of relevant publications. Once found, the full text is manually inspected to identify the gene product of interest, the relevant GO terms, and the evidence code to indicate the type of supporting evidence, e.g. mutant phenotype or genetic interaction, for inferring the relationship between a gene product and a GO term. Such a process is time-consuming and labor intensive, and thus many MODs are confronted with a daunting backlog of GO annotation. For instance, in recent years, TAIR's curation team has been able to curate only a fraction of newly published articles that contain information about Arabidopsis genes (<30%) (9). It is thus clear that the manual curation process requires computer assistance, and this is seen in a growing interest in, and need for semi- or fully automated curation pipelines for assisting biocuration (10-20). In particular, a number of studies (21-29) have attempted to (semi-)automatically predict GO terms from text including a previous BioCreative challenge task (30). However, few studies have proven to be useful with regard to assisting real world GO curation. Based on a recent study, enhanced text-mining capabilities to automatically recognize GO terms from full text remains one of the most in-demand tasks among the biocuration community (31).

As concluded in the previous BioCreative task (30,32), one of the main difficulties was “the lack of a high quality training set consisting in the annotation of relevant text passages”. Such a training set in practice provides the evidence for human curators to make associated GO annotations. To advance the development of automatic systems for GO curation, we propose to create a corpus that includes the GO evidence text along with three essential elements of GO annotations: 1) a gene or gene product, 2) a GO term (e.g., receptor-mediated endocytosis), and 3) a GO evidence code (e.g., Inferred from Mutant Phenotype (IMP)). The evidence texts for GO annotations may be derived from a single sentence, or multiple continuous, or discontinuous, sentences. The evidence for a GO annotation could also be derived from multiple lines of experimentation, leading to multiple text passages in a paper supporting the same annotation. Since many learning-based text-mining algorithms rely on both positive and negative training instances, it is important to be as thorough as possible when manually annotating sentences. It is therefore important to capture all of the curation-relevant sentences to ensure the positive and negative sets are as distinct as possible.

The exhaustive capture of evidence text in full-length articles makes our dataset, namely the BC4GO corpus, unique among the many previously annotated corpora (e.g.(33-36)) for the BioNLP research community. To our best knowledge, BC4GO is the only publicly available corpus that contains textual annotation of GO terms in accordance with the general practice of GO annotation (8) by professional GO curators. For instance, while in a previous study (17) every mention related to a GO concept was annotated, in BC4GO we have annotated only those GO terms that represent experimental findings in a given full-text paper.

Methods and Materials

Annotators

Through the BioCreative IV User Advisory Group, we recruited eight expert curators from five different MODs: FlyBase (2 curators), MaizeGDB (1 curator), RGD (3 curators), TAIR (1 curator), and WormBase (1 curator). All our curators are experienced in GO manual annotation.

Annotation Guidelines

For achieving consistent annotations between annotators, the task organizers followed the usual practice of corpus annotation (33-37): first we drafted a set of annotation guidelines and then asked each of our annotators to practice them on a shared article as part of the training process. The results of their annotations on the common article were shared among all annotators and subsequently the discrepancies in their annotations were discussed. Based on the discussion, the annotation guidelines were revised accordingly. For brevity, we only discuss below the two kinds of evidence text passages we chose to capture. The detailed guidelines are publicly available at the corpus download website: http://www.biocreative.org/resources/corpora/bc-iv-go-task-corpus/

1. Experiment Type: These sentences describe experimental results and can be used to make a complete GO annotation (i.e., the entity being annotated, GO term, and GO evidence code). The annotation of such sentences is required throughout the paper, including the abstract, and any supporting summary paragraphs such as ‘Author summary’ or ‘Conclusions’.

Ex1: *On the other hand, the amount of UNC-60B-GFP was reduced and UNC-60A-type mRNAs, UNC60A-RFP and UNC-60A-Experiment, were detected in asd-2 and sup-12 mutants (Figure 2H, lanes 2 and 3), consistent with their colour phenotypes shown in Figure 2C and 2A, respectively.* (PMC3469465)

This sentence contains information about:

The gene/protein entities: *asd-2* and *sup-12*
GO term: regulation of alternative mRNA splicing, via spliceosome (GO:0000381)
GO evidence code: Inferred from Mutant Phenotype (IMP)

2. Summary Type: Distinct from statements that describe the details of experimental findings, papers also include many statements that summarize these findings. These summary statements don’t necessarily indicate exactly *how* the information was discovered, but often contain concise language about *what* was discovered. Such sentences are helpful to capture because they may inform GO term selection in a concise manner despite the lack of information about evidence code selection.

Ex2: *Taken together, our results demonstrate that muscle-specific splicing factors ASD-2 and SUP-12 cooperatively promote muscle-specific processing of the unc-60 gene, and provide insight into the mechanisms of complex pre-mRNA processing; combinatorial regulation of a single splice site by two tissue-specific splicing regulators determines the binary fate of the entire transcript.* (PMC3469465)

The gene/protein entities: ASD-2 and SUP-12
GO term: regulation of alternative mRNA splicing, via spliceosome (GO:0000381)
GO evidence code: N/A

Article Selection

The 200 articles in the BC4GO corpus are chosen from annotators’ existing annotation workload at their respective MODs. Such a protocol minimizes the additional workload to our curators while at the same time guarantees the curated papers are representative of real-life GO annotations. Another requirement is that annotated articles are published in a list of select journals (e.g. PLoS Genetics) in PubMed Central (PMC) that allow free access and text analysis.

Annotation Tool

A web-based annotation tool was developed for use in the annotation process as shown below in Figure 1. The tool allows the upload of full text articles in either HTML or XML formats and subsequently displays the article in a Web browser. Currently, the tool allows the annotator to select and highlight a single sentence, or multiple sentences (regardless of whether they are contiguous or not) as GO evidence text. When a sentence is highlighted, a pop-up window appears for annotators to enter required GO annotation information: a GO term, a GO evidence code, and associated gene(s). The tool also allows the annotators to preview their annotations before committing them to the database. Annotation results of each paper can be downloaded as HTML files.

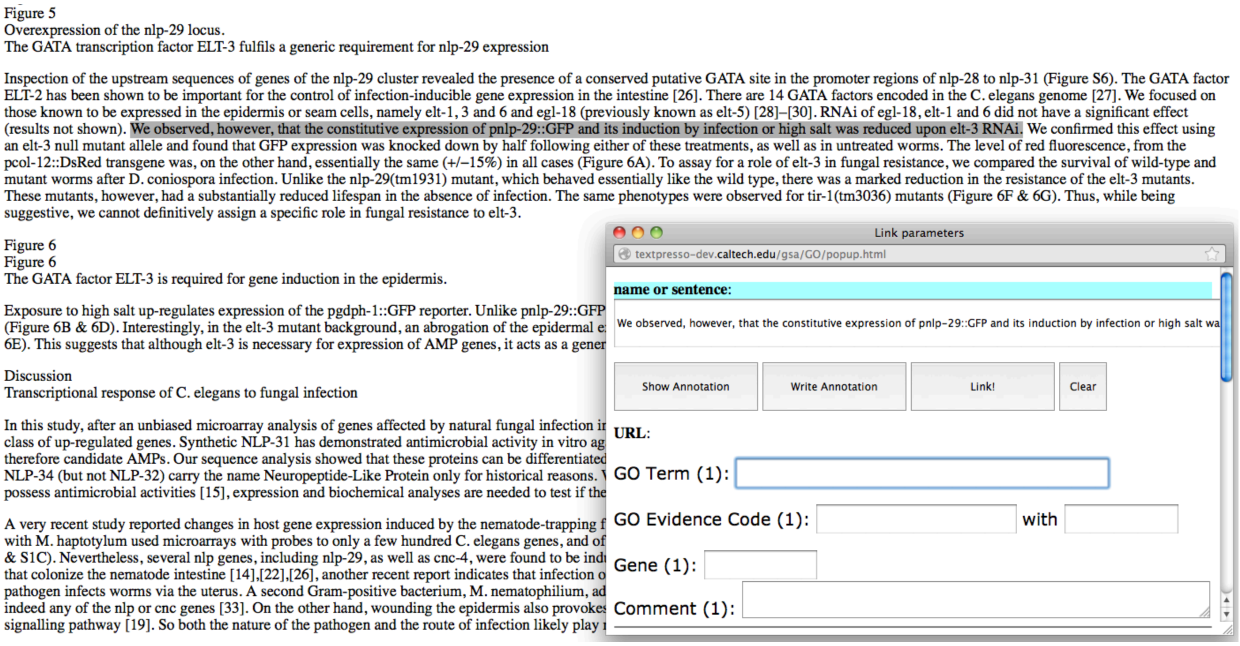


Figure 5
Overexpression of the nlp-29 locus.
The GATA transcription factor ELT-3 fulfils a generic requirement for nlp-29 expression

Inspection of the upstream sequences of genes of the nlp-29 cluster revealed the presence of a conserved putative GATA site in the promoter regions of nlp-28 to nlp-31 (Figure S6). The GATA factor ELT-2 has been shown to be important for the control of infection-inducible gene expression in the intestine [26]. There are 14 GATA factors encoded in the C. elegans genome [27]. We focused on those known to be expressed in the epidermis or seam cells, namely elt-1, 3 and 6 and egl-18 (previously known as elt-5) [28]–[30]. RNAi of egl-18, elt-1 and 6 did not have a significant effect (results not shown). We observed, however, that the constitutive expression of pnlp-29::GFP and its induction by infection or high salt was reduced upon elt-3 RNAi. We confirmed this effect using an elt-3 null mutant allele and found that GFP expression was knocked down by half following either of these treatments, as well as in untreated worms. The level of red fluorescence, from the pcol-12::DsRed transgene was, on the other hand, essentially the same (+/−15%) in all cases (Figure 6A). To assay for a role of elt-3 in fungal resistance, we compared the survival of wild-type and mutant worms after D. coniospora infection. Unlike the nlp-29(tm1931) mutant, which behaved essentially like the wild type, there was a marked reduction in the resistance of the elt-3 mutants. These mutants, however, had a substantially reduced lifespan in the absence of infection. The same phenotypes were observed for tir-1(tm3036) mutants (Figure 6F & 6G). Thus, while being suggestive, we cannot definitively assign a specific role in fungal resistance to elt-3.

Figure 6
Figure 6
The GATA factor ELT-3 is required for gene induction in the epidermis.

Exposure to high salt up-regulates expression of the pgdph-1::GFP reporter. Unlike pnlp-29::GFP
(Figure 6B & 6D). Interestingly, in the elt-3 mutant background, an abrogation of the epidermal e
6E). This suggests that although elt-3 is necessary for expression of AMP genes, it acts as a gener

Discussion
Transcriptional response of C. elegans to fungal infection

In this study, after an unbiased microarray analysis of genes affected by natural fungal infection ir
class of up-regulated genes. Synthetic NLP-31 has demonstrated antimicrobial activity in vitro ag
therefore candidate AMPs. Our sequence analysis showed that these proteins can be differentiated
NLP-34 (but not NLP-32) carry the name Neuropeptide-Like Protein only for historical reasons. V
possess antimicrobial activities [15], expression and biochemical analyses are needed to test if the

A very recent study reported changes in host gene expression induced by the nematode-trapping f
with M. haptotylum used microarrays with probes to only a few hundred C. elegans genes, and of
& S1C). Nevertheless, several nlp genes, including nlp-29, as well as cnc-4, were found to be indu
that colonize the nematode intestine [14],[22],[26], another recent report indicates that infection o
pathogen infects worms via the uterus. A second Gram-positive bacterium, M. nematophilium, ad
indeed any of the nlp or cnc genes [33]. On the other hand, wounding the epidermis also provokes
signalling pathway [19]. So both the nature of the pathogen and the route of infection likely play r

Link parameters

textpresso-dev.caltech.edu/gsa/GO/popup.html

name or sentence:

We observed, however, that the constitutive expression of pnlp-29::GFP and its induction by infection or high salt wa

URL:

GO Term (1):

GO Evidence Code (1): with

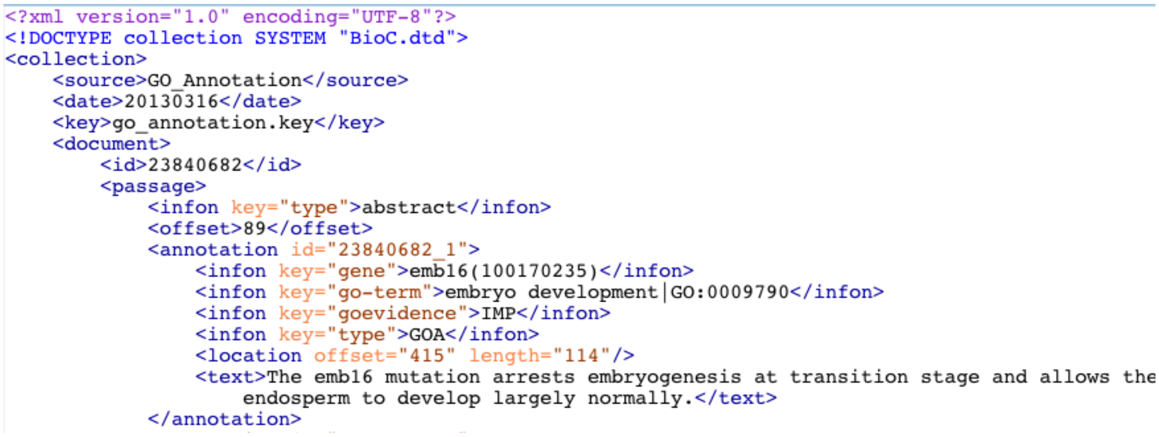
Gene (1):

Comment (1):

Figure 1. Screenshot of the annotation tool. When a line or more of text is highlighted, a pop-up window appears where annotation data is entered.

Final Data Dissemination

Both full-text articles and associated GO annotations (downloaded from PMC and the annotation tool, respectively) were further processed before releasing to the task participants. Specifically, we chose to format our data using the recently developed BioC standard for improved interoperability (38). First, for the 200 full-text articles, we converted their XMLs from the PMC format to the BioC format. Next, we extracted annotated sentences from downloaded HTML files and identified their offsets in the generated BioC XML files. Finally, for each article we created a corresponding BioC XML file for the associated GO annotations. Figure 2 shows a snapshot of our final released annotation files where one complete GO annotation is presented with the BioC format. For the gene entity, we provide both the gene mention as it appeared in the text and its corresponding NCBI Gene identifier.



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE collection SYSTEM "BioC.dtd">
<collection>
    <source>GO_Annotation</source>
    <date>20130316</date>
    <key>go_annotation.key</key>
    <document>
        <id>23840682</id>
        <passage>
            <infon key="type">abstract</infon>
            <offset>89</offset>
            <annotation id="23840682_1">
                <infon key="gene">emb16(100170235)</infon>
                <infon key="go-term">embryo development|GO:0009790</infon>
                <infon key="goevidence">IMP</infon>
                <infon key="type">GOA</infon>
                <location offset="415" length="114"/>
                <text>The emb16 mutation arrests embryogenesis at transition stage and allows the
                    endosperm to develop largely normally.</text>
            </annotation>
```

Figure 2. A sample of GO annotation in BioC format.

Results and Discussion

Corpus Statistics

The task participants are provided with three data datasets comprising a total of 200 full-text articles. Table 1 shows the number of articles curated by each MOD. On average, each curator contributed about 25 articles for the task during this time period.

Table 1. Number of curated articles per MOD.

Data Set	FlyBase	MaizeGDB	RGD	TAIR	WormBase	Total
Training Set	19	21	43	10	7	100
Development Set	8	5	25	4	8	50
Test Set	12	4	20	7	7	50
Subtotal per team	39	30	88	21	22	200

Table 2 shows the main characteristics of the BC4GO corpus. Each annotation includes four elements: the gene/protein entity, GO term, GO evidence code, and evidence text (See **Figure**). Note that one text passage can often provide evidence for annotating more than one gene, as well as more than one GO term. Therefore, we show in the last column of Table 2 the counts of evidence text passages in three different ways. The first number shows that the total number of text passages with respect to GO annotations: Over 5,000 text passages were used in the annotation of 1,311 unique GO terms. So on average, each GO term is associated with four different evidence text passages in our corpus. The second number (5,162) shows the total number of text passages with respect to different genes: For each of the 665 unique genes in our corpus, there are about 7.8 associated text passages. Finally, the last number is the total number of unique text passages annotated in our corpus regardless of their association to either gene or GO terms.

Table 2. Overall statistics of the annotated corpus.

Data Set	Articles	Genes (unique)	GO terms (unique)	Evidence text passages w.r.t. GO/Gene/Unique
Training Set	100	300	566	2,213/2,234/1,704
Development Set	50	171	367	1,299/1,247/963
Test Set	50	194	378	1,763/1,681/1,253
Total	200	665	1,311	5,275/5,162/3,920

From Table 2, we can compute that the average number of genes annotated in each article is 3.3, and the average number GO terms associated with each gene is 2.0 in our corpus. Furthermore, as mentioned before, we have annotated two types of evidence text, depending on whether they contain experimental information or not. Accordingly, the two kinds are distinguished in our annotations by the presence or absence of associated evidence code. For the total 3,920 unique pieces of evidence text, the majority (~70%) of them contain experimental evidence.

The location of evidence text in the paper

Figure 3 shows the proportion of all evidence text in different parts of the article. As can be seen, the most informative location for extracting GO evidence text is the Results section, followed by the Discussion Section. Some GO evidence text also appears in the Table or Figure legend. Within the full text article, the Introduction/Background and Methods sections contain the least amount of information for complete GO annotation. Figure 3 also shows the limitation of using article abstracts for GO annotation: only 11.65% of the annotated text is found in the Title and Abstract combined. This finding further confirms the importance of using full text for GO annotation.

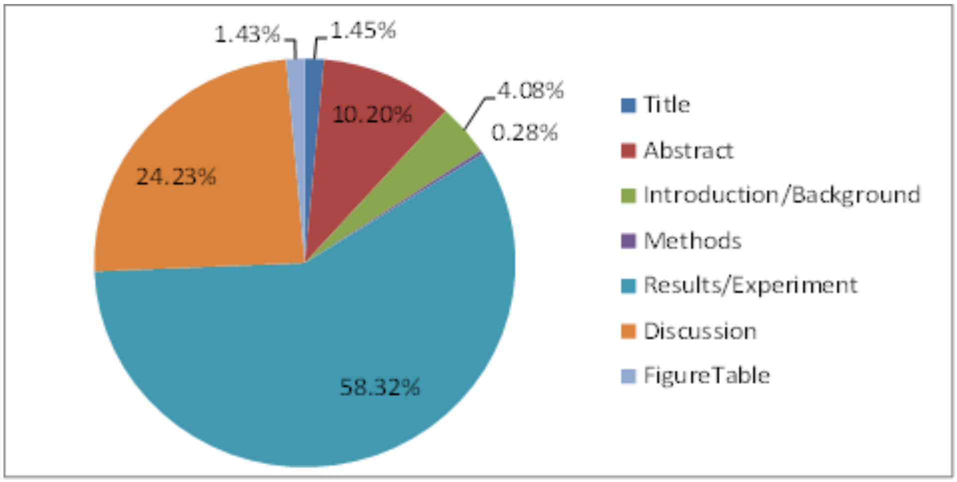


Figure 3. The proportion of annotated evidence text in different parts of the article.

Conclusions and Future Work

Through collaboration with professional GO curators from five different MODs, we created a corpus for the development and evaluation of automated methods for identifying GO terms from full-text articles. The resulting BC4GO corpus is large-scale and the only one of its kind. We expect our BC4GO corpus to become a valuable resource for the BioNLP research community. We hope to see improved performance and accuracy of text mining for GO terms through the use of our annotated corpus in the BioCreative IV GO task and beyond.

There are several limitations of this work that warrant further investigation. First, in order to ensure the positive and negative sentences are as distinct as possible, we asked our annotators to mark up every occurrence of GO evidence text. As a result, it greatly increased the annotation workload for each individual annotator. Meanwhile, to maximize the number of annotated articles, we chose to assign one annotator per article. In other words, our articles are not double annotated. Second, despite all our best efforts in ensuring consistent annotations (e.g. creating annotation guidelines, and providing annotator training), there will always be variation in the depth of annotation between curators and organisms. For instance, there may be gray areas where some curators will select a sentence relating to a phenotype as a GO sentence, while others do not. In the future, we plan to assess the inter-annotator agreement for our corpus.

Authors' Contributions:

Conceived and designed the annotation experiment: ZL, KVA, DL, CNA. Developed the annotation guidelines: ZL, KVA, PM, DL, ST. Developed the annotation tool: JD, KVA, HMM, PWS. Performed the annotation experiment: MLS, PM, SJFL, KVA, DL, SJW, GTH, ST, CNA. Analyzed the annotated data: YM, CW, ZL. Wrote the paper: ZL. All authors read and approved the final manuscript.

Acknowledgments

We would like to thank Don Comeau, Rezarta Dogan and John Wilbur for general discussion and technical assistance in using BioC, and in particular to Don Comeau for providing us source PMC articles in the BioC XML format. We also thank Lynette Hirschman, Cathy Wu, Kevin Cohen, Martin Krallinger, and Thomas Wiegers from the BioCreative IV organizing committee for their support, and Judith Blake, Andrew Chatr-aryamontri, Sherri Matis, Fiona McCarthy, Sandra Orchard, and Phoebe Roberts from the BioCreative IV User Advisory Group for their helpful discussions. This research is supported by the Intramural Research Program of the NIH, National Library of Medicine (CW, YM, & ZL), by the USDA ARS (MLS), by grants from the National Human Genome Research Institute at the US National Institutes of Health # HG004090, # HG002223 and # HG002273, and by National Science Foundation grant ABI-1062520.

References

1. Harris, M.A., Clark, J., Ireland, A.*, et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, **32**, D258-261.
2. Balakrishnan, R., Harris, M.A., Huntley, R.*, et al.* (2013) A guide to best practices for Gene Ontology (GO) manual annotation. *Database : the journal of biological databases and curation*, **2013**, bat054.
3. Hill, D.P., Berardini, T.Z., Howe, D.G.*, et al.* (2010) Representing ontogeny through ontology: a developmental biologist's guide to the gene ontology. *Mol Reprod Dev*, **77**, 314-329.
4. Mutowo-Meullenet, P., Huntley, R.P., Dimmer, E.C.*, et al.* (2013) Use of Gene Ontology Annotation to understand the peroxisome proteome in humans. *Database : the journal of biological databases and curation*, **2013**, bas062.
5. Ochs, M.F., Peterson, A.J., Kossenkov, A.*, et al.* (2007) Incorporation of gene ontology annotations to enhance microarray data analysis. *Methods Mol Biol*, **377**, 243-254.
6. Burge, S., Kelly, E., Lonsdale, D.*, et al.* (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database : the journal of biological databases and curation*, **2012**, bar068.
7. Barrell, D., Dimmer, E., Huntley, R.P.*, et al.* (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res*, **37**, D396-403.
8. Balakrishnan, R., Harris, M.A., Huntley, R.*, et al.* (2013) A guide to best practices for Gene Ontology (GO) manual annotation. *Database (oxford)*, **2013**, bat054.
9. Li, D., Berardini, T.Z., Muller, R.J.*, et al.* (2012) Building an efficient curation workflow for the Arabidopsis literature corpus. *Database (Oxford)*, **2012**.
10. Aerts, S., Haeussler, M., van Vooren, S.*, et al.* (2008) Text-mining assisted regulatory annotation. *Genome Biol*, **9**, R31.
11. Arighi, C.N., Carterette, B., Cohen, K.B.*, et al.* (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database : the journal of biological databases and curation*, **2013**, bas056.
12. Arighi, C.N., Lu, Z., Krallinger, M.*, et al.* (2011) Overview of the BioCreative III Workshop. *BMC Bioinformatics*, **12 Suppl 8**, S1.
13. Li, D., Berardini, T.Z., Muller, R.J.*, et al.* (2012) Building an efficient curation workflow for the Arabidopsis literature corpus. *Database : the journal of biological databases and curation*, **2012**, bas047.
14. Neveol, A., Wilbur, W.J., Lu, Z. (2012) Improving links between literature and biological data with text mining: a case study with GEO, PDB and MEDLINE. *Database : the journal of biological databases and curation*, **2012**, bas026.
15. Van Auken, K., Jaffery, J., Chan, J.*, et al.* (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*, **10**, 228.
16. Wu, C.H., Arighi, C.N., Cohen, K.B.*, et al.* (2012) BioCreative-2012 virtual issue. *Database : the journal of biological databases and curation*, **2012**, bas049.
17. Wei, C.-H., Harris, B.R., Li, D.*, et al.* (2012) Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database(oxford)*, bas041.

18. Wei, C.-H., Kao, H.-Y., Lu, Z. (2012) PubTator: A PubMed-like interactive curation system for document triage and literature curation. *Proceedings of the BioCreative 2012 workshop*, Washington, D.C., pp. 20-24.
19. Wei, C.-H., Kao, H.-Y., Lu, Z. (2013) PubTator: a Web-based text mining tool for assisting Biocuration. *Nucleic Acids Res*, **41**, W518-W522.
20. Neveol, A., Wilbur, W.J., Lu, Z. (2011) Extraction of data deposition statements from the literature: a method for automatically tracking research results. *Bioinformatics*, **27**, 3306-3312.
21. Raychaudhuri, S., Chang, J.T., Sutphin, P.D*., et al.* (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research*, **12**, 203–214.
22. Daraselia, N., Yuryev, A., Egorov, S*., et al.* (2007) Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. *BMC Bioinformatics*, **8**, 243.
23. Auken, K.V., Jaffery, J., Chan, J*., et al.* (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. *BMC Bioinformatics*, **10**, 228.
24. Costanzo, M.C., Park, J., Balakrishnan, R*., et al.* (2011) Using computational predictions to improve literature-based Gene Ontology annotations: a feasibility study. *Database (Oxford)*, **2011**, bar004.
25. Park, J., Costanzo, M.C., Balakrishnan, R*., et al.* (2012) CvManGO, a method for leveraging computational predictions to improve literature-based Gene Ontology annotations. *Database (Oxford)* **2012**, bas001.
26. Rak, R., Rowley, A., Black, W*., et al.* (2012) Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database (Oxford)*, **2012**, bas010.
27. Gobeill, J., Pasche, E., Vishnyakova, D*., et al.* (2013) Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases. *Database (oxfored)*, **2013**, bat041.
28. Koike, A., Niwa, Y., Takagi, T. (2004) Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, **2005**, 7.
29. Cakmak, A., Ozsoyoglu, G. (2008) Discovering gene annotations in biomedical text databases. *BMC Bioinformatics*, **9**, 143.
30. Blaschke, C., Leon, E.A., Krallinger, M*., et al.* (2005) Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, **6**, S16.
31. Lu, Z., Hirschman, L. (2012) Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database (Oxford)*, **2012**, bas043.
32. Camon, E.B., Barrell, D.G., Dimmer, E.C*., et al.* (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, **6 Suppl 1**, S17.
33. Bada, M., Eckert, M., Evans, D*., et al.* (2012) Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, **13**, 161.
34. Kim, J.D., Ohta, T., Tateisi, Y*., et al.* (2003) GENIA corpus--semantically annotated corpus for bio-textmining. *Bioinformatics*, **19 Suppl 1**, i180-182.
35. Dogan, R.I., Lu, Z. (2012) An improved corpus of disease mentions in PubMed citations. *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Montreal, Canada, pp. 91-99.

36. Smith, L., Tanabe, L.K., Ando, R.J.*, et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome Biol*, **9 Suppl 2**, S2.
37. Lu, Z., Bada, M., Ogren, P.V.*, et al.* (2006) Improving biomedical corpus annotation guidelines. *Proceedings of the Joint BioLINK and 9th Bio-Ontologies Meeting*, Fortaleza, Brazil, pp. 89-92.
38. Comeau, D.C., Islamaj Dogan, R., Ciccarese, P.*, et al.* (2013) BioC: a minimalist approach to interoperability for biomedical text processing. *Database : the journal of biological databases and curation*, **2013**, bat064.